Knowledge Discovery in Clinical Databases based on Variable Precision Rough Set Model

Shusaku Tsumoto*1, Wojciech Ziarko*2, Ning Shan*2, and Hiroshi Tanaka*1
*1 Department of Information Medicine,
Medical Research Institute, Tokyo Medical and Dental University
1-5-45 Yushima, Bunkyo-ku Tokyo 113 Japan
*2 Department of Computer Science, University of Regina
Regina, Saskatchewan, Canada S4S 0A2

ABSTRACT

Since a large amount of clinical data are being stored electronically, discovery of knowledge from such clinical databases is one of the important growing research area in medical informatics. For this purpose, we develop KDD-R(a system for Knowledge Discovery in Databases using Rough sets), an experimental system for knowledge discovery and machine learning research using variable precision rough sets (VPRS) model, which is an extension of original rough set model. This system works in the following steps. First, it preprocesses databases and translates continous data into discretized ones. Second, KDD-R checks dependencies between attributes and reduces spurious data. Third, the system computes rules from reduced databases. Finally, fourth, it evaluates decision making. For evaluation, this system is applied to a clinical database of meningoencephalitis, whose computational results show that several new findings are obtained.

1. INTRODUCTION

Knowledge discovery in clinical databases is an important research area in medical informatics. Most of medical data, such as patient records, laboratory data, are now being stored electronically, and the amount of clinical databases will be too huge, so that even medical experts cannot deal with such large databases. Thus, a computer-based approach is promising to solve this difficult situation.

In this paper, we introduce a system KDD-R(a system for Knowledge Discovery in Databases using Rough sets), based on Variable Precision Rough Set(VPRS) model[4].

This system works as follows. First, it preprocesses databases and translates continuous data into discretized ones. Second, KDD-R checks dependencies between attributes and reduces spurious data. Third, the system computes rules from reduced databases. Finally, fourth, it evaluates decision making.

For evaluation, we apply KDD-R to a clinical database of meningoencephalitis, whose computational results show that several new findings are obtained from clinical databases.

The paper is organized as follows: Section 2 gives an overview of KDD-R system. Then, Section 3 to 5 shows the main computational units of KDD-R. Next, Section 6 presents experimental results. Finally, Section 7 concludes this paper.

Due to the limitation of space, we do not fully discuss the concepts of rough sets and variable precision model, although the main concepts needed are shown in each section. For further information, readers could refer to [1,2,3,4,7,8].

2. KDD-R and PROBLEM REPRESENTATION

KDD-R is an open tool-box, implemented in C under UNIX. Currently, it is menu driven with X-Windows interface implementation in the works. The system contains, among others, the following primary functional units:

(1) Data preprocessing unit, responsible for mapping original data set into discretized form, either by using user-supplied discretization formula or by applying an automatic discretization algorithm.

(2) A unit for analysis of dependencies among attributes and elimination of superfluous attributes. The unit also enables one to compute generalized attribute reducts and cores[1], as defined in the VPRS model.

(3) A unit for computation of rules from data. This unit computes all, or some, approximate rules with decision probabilities, where the probabilities are restricted by lower and upper limit parameters[2] specifying the area of user interest. The rules can be computed for a selected reduct using the method of decision matrix[6]. The unit can also be used to compute maximal approximate rules, that is the maximal elements

[1]Reducts denotes minimal number of independent attributes, and cores is derived as intersection of all the reducts.

in the partial ordering of rules with respect to relation of inclusion of data sets supporting each rule. Such rules are the strongest in terms of available data support and independent from each other. The user can also select among all rules or the rules forming minimal covering of the target concept.

(4) Decision unit which can be used in situations requiring system's advice based on previously accumulated information. The decision unit is combining all the available evidence, expressed in the form of rules, to suggest the most likely decision for a given situation.

The object of the analysis in KDD-R is a flat relational table with rows containing information about objects or situations from a certain universe of discourse, expressed in terms of attribute values. The permissible attribute values are integers and reals.

In the analysis of the relationship, each value v of the decision attribute is treated as a set $|v|$ of table rows matching this value. Our primary problem is to produce, in automatic or semi-automatic fashion, plausible and strongly supported by available evidence hypotheses about the true nature of the relationship between the occurrence of combinations of some properties of condition attributes values in real objects and the occurrence of the value v of the decision attribute. In doing that, we construct an associated secondary table T_v for each decision attribute value of interest, with unchanged condition attributes and a new binary decision attribute corresponding to the characteristic function of the set $|v|$. Following this step, the original problem with m decision attribute values is decomposed by KDD-R into m subproblems, each with a binary decision attribute.

3. DATA PROCESSING

Prior to running KDD-R, the user is required to provide several control parameters whose definitions will be introduced gradually throughout the paper. The preprocessing unit of the system converts each constructed table T_v with binary decision attributes[2] into a corresponding table with all condition attributes discretized. One of the parameters supplied by the user is the number n of discrete condition attributes in the preprocessed table.

Data preprocessing involves defining a secondary set of features which are functions of original attribute values. The original attribute values are often to detailed to capture repetitive regularities, or patterns occurring in the data. The secondary feature definitions can be either provided by the user based on domain knowledge, or can be produced automatically using, for example, some statistical techniques[2].

[2]Conditional attributes correspond to the premise of a proposition, and decision attributes are equivalent to the conclusion of a proposition.

KDD-R enables the user to define his/her own discrete secondary features in terms of properly selected value ranges. The other possibility is to get the system to generate the definition of secondary features. This option applies only to real-valued attributes. In this process, the real-valued attributes are replaced by one, or more three-valued discrete attributes corresponding to value ranges. For a given range $(a, b>$, the new value $v_{(a,b>}$ assigned to value v is given by

$$v_{(a,b>} = \begin{cases} 0 & \text{if } v \leq a \\ 1 & \text{if } a \leq v \leq b \\ 2 & \text{if } b < v. \end{cases}$$

When constructing the three-valued representation for condition attributes of each of m data tables with binary decision attributes, the system is performing internal search looking for ranges which maximize the given range quality criterion $Q(r_A)$.

To describe the criterion $Q(r_A)$, let $|D = 1|$ denote the set of data rows with the secondary decision attribute value equal to 1, and let $|r_A|$ be the set of rows with values of the attribute A falling into the range $r_A = (a, b>$. The criterion of range quality $Q(r_A)$ is based on the estimation of the following conditional probabilities:

(1) $P(r_A|D = 1)$, the probability that an object has the value of attribute A falling in the range r_A, provided that the value of the decision attribute is 1;

(2) $P(D = 1|r_A)$, the probability that an object has the value of the decision attribute given by $D = 1$ if the value of A attribute belongs to the range r_A;

(3) $P(r_A|D = 0)$, the probability that an object has the value of attribute A falling in the range r_A if the value of the secondary decision attribute is $D = 0$;

(4) $P(D = 0|r_A)$, the probability that the value of the decision attribute is $D = 0$ if an object has the value of the condition attribute belonging to r_A.

Intuitively, if the random event $(v \in r_A)$ and $(D = 1)$ are well connected in statistical sense then both measures (1) and (2) should yield high values and the measures (3) and (4) should yield low values. Consequently, based on this intuition the range quality measure used by KDD-R is

$$\begin{aligned} Q(r_A) &= P(r_A|D=1) + P(D=1|r_A) \\ &\quad - P(r_A|D=0) - P(D=0|r_A). \end{aligned}$$

Clearly, $-2 \leq Q(r_A) \leq 2$. $Q(r_A)$ can be seen as a measure of bias of the set $|r_A|$ towards the set $|D = 1|$

with two extremes: $Q(r_A) = 2$ if $|r_A| = |D = 1|$ and $Q(r_A) = -2$ if $|r_A| = |D = 0|$.

Following the range search process, m secondary tables, each with n best three-valued condition attributes, are created and passed to the further stages of analysis.

4. VPRS-BASED DATA ANALYSIS

Following the quantization stage, each one of the m tables can be analyzed to investigate the relationship between the occurrence of the value $D = 1$ of the secondary decision attribute and the discrete values of n condition attributes. The theoretical model behind the analytical routines used in this process is the Variable Precision Rough Set (VPRS) model[4]. In comparison to the original Pawlak's definition [1], the VPRS technique provides a degree of flexibility in specifying the lower bound, boundary region and the negative region of a set Y by allowing a controlled degree of overlap of lower bound atoms and negative region atoms with the set Y complement or the set Y, respectively.

More precisely, for given lower and upper limit parameters ℓ and u respectively, such that $0 \leq \ell \leq u \leq 1$, the ℓ-lower approximation (or ℓ-positive region) of the subset $Y \subseteq U$ in the approximation space $A = (U, R)$ is given by

$$\underline{R}_\ell(Y) = \bigcup\{E \in R^* : c(E, Y) \leq \ell\},$$

where R^* is a collection of the classes of abstraction of the equivalence relation $R \subset U \times U$ and $c(E, Y)$ is a classification factor, or an apparent error rate defined as

$$c(E, Y) = 1 - \frac{card(E \cap Y)}{card(E)}.$$

The classification factor is a measure of the relative degree of intersection of an atom E with the complement of set Y. The equivalence relation $R \subset U \times U$ in KDD-R is treating any two rows of a secondary table T_v with identical values of secondary condition attributes as equivalent.

The (ℓ, u)-boundary region of the set Y is given by

$$BNR_{\ell,u}(Y) = \bigcup\{E \in R^* : \ell < c(E, Y) < u\},$$

and the u-negative region is defined as

$$NEG_u(Y) = \bigcup\{E \in R^* : c(E, Y) \geq u\}.$$

Before using KDD-R, when specifying system parameters, user is asked to provide the lower and upper limit parameters, ℓ and u. He or she also needs to indicate whether the data analysis will be focused on the ℓ-lower bound or on the u-upper bound of each value of the decision attribute, where the u-upper bound is simply a union of ℓ-lower bound and (ℓ, u)-boundary region. For the sake of simplicity, we will assume here that the analysis will be focused on the ℓ-lower bound. This means that KDD-R will compute the ℓ-lower bounds $\underline{R}_\ell(|D = 1|)$ of sets $|D = 1|$ in all m tables with binary decision attributes. Following this computation, the KDD-R user has the following choice of options to analyze the relationship between the discretized condition attributes and the approximation regions of $|D = 1|$.

a. Computing the measure of dependency between discrete condition attributes and the secondary decision attribute, $DEP(C, D)$, as given by the expression

$$\frac{card(\underline{R}_\ell(|D = 1|)) + card(NEG_u(|D = 1|))}{card(U)},$$

where C is a set of condition attributes used to obtain the classification of objects into identity classes corresponding to the relation R.

b. Computing the degree of accuracy of $|D = 1|$ approximation using two measures:

$$M_1(D = 1) = \frac{card(\underline{R}_\ell(|D = 1|) \cap |D = 1|)}{card(|D = 1|)}$$

and

$$M_2(D = 1) = \frac{card(\underline{R}_\ell(|D = 1|))}{card(\overline{R}_u(|D = 1|))}.$$

c. Computing one relative reduct[1] or all relative reducts of condition attributes with respect to preservation of both ℓ-lower bound and u-negative region of $|D = 1|$.

d. Computing one relative reduct or all relative reducts of condition attributes with respect to preservation of ℓ-lower bound (or u-upper bound, if selected by the user) of $|D = 1|$.

e. Computing core attributes[1] with respect to the given dependency function.

The computation of a single relative reduct either requires the user to provide a priority ordering on the condition attributes or range quality measure, as defined in Section 3, is used by the system to produce the priority ordering of secondary attributes. The computation involves testing each secondary attribute, in the reverse order of priority, by removing it from the table and checking whether the dependency with the decision attribute is changed. If the dependency was not affected by the removed attribute, the attribute is eliminated permanently, otherwise it is returned to the table. The algorithm for computation of all relative reducts accomplished with the help of the decision matrix method, as implemented in KDD-R, is fully described in [8].

5. COMPUTATION OF RULES

Computation of rules, besides computation of reducts, is one of the most important activities carried out by KDD-R. Similar to computation of reducts, the decision matrix technique[6] is used here to find all minimal length rules for ℓ-lower bound (or u-upper bound) of the set $|D = 1|$ of each of the discretized secondary tables. The technique involves finding prime implicants, which correspond to maximally general rules, described as a Boolean function called decision function[5,6]. The decision matrix method, as applied to computation of rules, is briefly summarized below.

Before we define the notion of a decision matrix, we will assume some notational conventions. We will assume that all classes E_i of R^* such that $E_i \subseteq \underline{R}_\ell(|D = 1|)$ and all classes E_j such that $E_j \subseteq NEG_u(|D = 1|)$ are separately numbered with subscripts i $(i = 1, 2, ...\gamma)$ and j $(j = 1, 2, ...\rho)$ respectively.

A decision matrix $M = (M_{ij})_{\gamma \times \rho}$ with respect to ℓ-lower approximation of $|D = 1|$ (or u-upper approximation of $|D = 1|$) is defined by

$$M_{ij} = \{(A, A(E_i)) : A(E_i) \neq A(E_j)\},$$

where A is a secondary discretized condition attribute and $A(E_i)$ is the value of this attribute common to all objects belonging to the atom E_i.

The matrix entry M_{ij} contains all attribute-value pairs $(attribute, value)$ whose values are not identical on atoms E_i and E_j. M_{ij} represents the complete information distinguishing atomic classes E_i from E_j.

The set of decision rules computed for a given class E_i $(i = 1, 2, ...\gamma)$ is obtained by treating each element of M_{ij} as a propositional variable and forming a Boolean function (decision function)

$$B_i = \bigwedge_j \bigvee M_{ij},$$

where $\bigwedge$ and $\bigvee$ are respectively generalized conjunction and disjunction operators.

The prime implicants of the decision function correspond to maximally general rules for the ℓ-lower bound $\underline{R}_\ell(|D = 1|)$. By finding the prime implicants of all decision functions B_i $(i = 1, 2, ..., \gamma)$, all maximally general rules can be computed for the ℓ-lower bound of $|D = 1|$. Similarly, all rules can be found for the u-upper bound or the u-negative region of $|D = 1|$.

In addition to computing all rules for the ℓ-lower bounds, KDD-R has also the option of finding a subset of "best" rules. The "best" rules, or the maximal rules are the maximal elements in partial ordering of the rules with respect to the relationship of inclusion among their support sets. The support set of a rule, denoted as $supp(r)$, is defined here as a collection of rows of the original table satisfying the condition of the rule's condition part.

The identifiers of rows belonging to rule support set are shown for each rule computed by KDD-R. The rules with larger support sets are considered to be "stronger" and better asserted by the available evidence. Also, no maximal rule is covered by any other rule, so in this sense the maximal rules are also independent.

Another option available to KDD-R user is the possibility of finding the minimum, or close to minimum, subset of strongest rules covering the ℓ-lower approximation of $|D = 1|$. Since the basic rough sets model implemented in KDD-R is VPRS, the rules can be nondeterministic which means that more than one outcome is possible based on some rules. Because each range of values of an attribute A, r_A can be perceived as a nondeterministic, in general, rule $r_A \to (D = 1)$, it follows that all measures used for range evaluation are also applicable to rules. Consequently, when presenting each rule to the user, the system also provides estimates of conditional probabilities, with respect to all possible outcomes, and rule quality measure, exactly in the same way as they are defined in Section 3 for evaluation of ranges.

6. EXPERIMENTAL RESULTS

We apply KDD-R to a clinical database of meningoencephalitis collected from Matsudo Municipal Hospital in Japan. This database, described by 26 conditional attributes and one decision attribute(diagnosis of neurologists), has 96 training samples, composed of 66 viral cases and 30 bacterial ones. Using this database, we analyze what factors are important for differential diagnosis between viral and bacterial infection.

For viral meningitis, 15 positive and 18 negative region rules are derived. The best two or three rules for each region, which cover many training samples, are shown in the following:

Positive Region rules for (D:Viral Meningitis):

(1) Premise:
$(37.0 < BT \leq 39.0)\&(200 < CSF_{CELL} \leq 1000)$
Rule Coverage: $C(D) = 19$, $C(-D) = 0$,
Range Quality Measure: $Q(r) = 1.288$.

(2) Premise:
$(SEX = F)\&(200 < CSF_{CELL} \leq 1000)$
Rule Coverage: $C(D) = 21$, $C(-D) = 0$,
Range Quality Measure: $Q(r) = 1.318$.

Negative Region rules for (D:Viral Meningitis):

(1) Premise:
$(39.0 < BT \leq 40.2)\&(1000 < CSF_{CELL} \leq 63350)$
Rule Coverage: $C(D) = 0$, $C(-D) = 8$,
Range Quality Measure: $Q(r) = -1.242$.

(2) Premise:
$(49 < AGE \leq 59)\&(1000 < CSF_{CELL} \leq 63350)$
Rule Coverage: $C(D) = 0$, $C(-D) = 6$,
Range Quality Measure: $Q(r) = -1.182$.

Positive Region rules for (D:Bacterial Meningitis):

(1) Premise:
$(1.2 < CRP \leq 31.0)\&(1000 < CSF_{CELL} \leq 63350)$
Rule Coverage: $C(D) = 10$, $C(-D) = 0$,
Range Quality Measure: $Q(r) = 1.333$.

(2) Premise:
$(39.0 < BT \leq 40.2)\&(1000 < CSF_{CELL} \leq 63350)$
Rule Coverage: $C(D) = 8$, $C(-D) = 0$,
Range Quality Measure: $Q(r) = 1.267$.

Negative Region rules for (D:Bacterial Meningitis:

(1) Premise:
$(SEX = F)\&(200 < CSF_{CELL} \leq 1000)$
Rule Coverage: $C(D) = 0$, $C(-D) = 21$,
Range Quality Measure: $Q(r) = -1.304$.

(2) Premise:
$(0.0 \leq CRP \leq 0.6)\&(5 < CSF_{CELL} \leq 200)$
Rule Coverage: $C(D) = 0$, $C(-D) = 20$,
Range Quality Measure: $Q(r) = -1.290$.

In the above rules, BT and CSF_{CELL} denote body temperature and cell count in cerebulospinal fluid, respectively.

Each of the above rules is given by specifying its condition in the first row, followed by rule coverage (denoted by $C(.)$)[3] and rule quality measure $Q(r)$.

These obtained rules give us four interesting results. First, CSF_{cell} is the most important attribute-value pair, which corresponds to medical knowledge. Second, values of the markers for viral infection, such as BT, CRP, and CSF_{cell} are much lower than ones for bacterial infection. Third, interestingly, women do not often suffer from bacterial infection, compared with men. However, in medical context, such sex relationship has not been discussed[9]. Thus, this sexual relation seems to be dependent on our training samples. Examined databases clearly, it is found that most of the above patients suffers from chronic diseases, such as DM, LC, and sinusitis, which are the risk factors of bacterial meningitis. Fourth, age is also an important factor not to suspect viral meningitis, which also matches the fact that most old people suffers from chronic diseases. Interestingly, the first negative region rule shares only one case with the second negative one, which weakly suggests that $39.0 < BT \leq 40.2$ and $49 < AGE \leq 59$ be independent.

[3]Rule coverage is equal to the number of training samples which satisfies the premise of a given rule.

7. CONCLUSION

In this paper, we introduce a system KDD-R in order to discover knowledge in clinical databases. This system is a new software environment designed from the bottom up with the sole objective of providing a collection of rough sets-based tools for comprehensive data analysis and knowledge discovery using VPRS model. We apply KDD-R to a clinical database on meningoencephalitis. The results show that several new findings are obtained from clinical databases.

Acknowledgments

The authors would like to thank Dr. Kitano and Dr. Komatsu for a clinical database of meningitis. The research reported in this paper was supported in part by an operating grant from the Natural Sciences and Engineering Research Council of Canada.

References

1.Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning About Data. Dordrecht: Kluwer. 1991.
2.Ziarko W. (ed.) Rough Sets, Fuzzy Sets and Knowledge Discovery. London: Springer-Verlag. 1994.
3.Piatetsky-Shapiro G. and Frawley WJ. (eds.) Knowledge Discovery in Databases. Cambridge: MIT Press. 1991.
4.Ziarko W. Variable Precision Rough Set Model. Journal of Computer and System Sciences. 1993; 46:39-59.
5.Ziarko W. and Shan N. A Rough Set-Based Method for Computing All Minimal Deterministic Rules in Attribute-Value Systems, Computational Intelligence. 1995; 11, to appear.
6.Skowron A. and Rauszer C. The Discerniblity Matrices and Functions in Information Systems, In Slowinski, R. (ed.) Intelligent Decision Support: Handbook of Applications and Advances of Rough Sets Theory, Dordrecht: Kluwer. 1992.
7.Grzymala-Busse JW. LERS - a System Learning from Examples Based on Rough Sets, In Slowinski, R. (ed.) Intelligent Decision Support: Handbook of Applications and Advances of Rough Sets Theory. Dordrecht: Kluwer. 1992.
8.Shan N. and Ziarko W. Data-Based Acquisition and Incremental Modification of Classification Rules, Computational Intelligence. 1995; 11, (to appear).
9.Adams RD. and Victor M. *Principles of Neurology*, 5th edition, New York, McGraw-Hill. 1993.